INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Audiovisual recalibration of vowel categories

Matthias K. Franken[1,2], *Frank Eisner*[1], *Jan-Mathijs Schoffelen*[1], *Daniel J. Acheson*[1,2], *Peter Hagoort*[1,2], *James M. McQueen*[1,2]

[1]Radboud University, Donders Institute for Brain, Cognition and Behaviour, the Netherlands
[2]Max Planck Institute for Psycholinguistics, the Netherlands

m.franken@donders.ru.nl

Abstract

One of the most daunting tasks of a listener is to map a continuous auditory stream onto known speech sound categories and lexical items. A major issue with this mapping problem is the variability in the acoustic realizations of sound categories, both within and across speakers. Past research has suggested listeners may use visual information (e.g., lip-reading) to calibrate these speech categories to the current speaker. Previous studies have focused on audiovisual recalibration of consonant categories. The present study explores whether vowel categorization, which is known to show less sharply defined category boundaries, also benefit from visual cues.

Participants were exposed to videos of a speaker pronouncing one out of two vowels, paired with audio that was ambiguous between the two vowels. After exposure, it was found that participants had recalibrated their vowel categories. In addition, individual variability in audiovisual recalibration is discussed. It is suggested that listeners' category sharpness may be related to the weight they assign to visual information in audiovisual speech perception. Specifically, listeners with less sharp categories assign more weight to visual information during audiovisual speech recognition.

Index Terms: speech perception, categorical perception, perceptual learning

1. Introduction

Speech perception is a remarkably complex skill. One of the most obvious issues every listener has to deal with is the enormous amount of variability in the speech signal. Acoustic variability in the speech signal is due to various factors, including the phonological context, the speaker's mood, speaker idiosyncrasies, the speaker's accent or dialect, etc.

One way in which listeners can deal with this variability in the acoustic signal is by recalibrating speech sound categories using additional sources of information [1]. For example, the so-called Ganong effect shows that listeners may use lexical information to bias speech perception [2]. A number of studies exposed listeners to a series of words where one consonant was replaced with an ambiguous sound. For example, Dutch listeners were presented with words like <witlo?>, where <?> represents a sound ambiguous between /f/ and /s/, making the word ambiguous between <witlof> ("chicory") and <witlos> (a pseudoword). The results showed that after exposure to a series of these items (where an /f/ interpretation would yield a known lexical item while the alternative /s/ interpretation would yield a pseudoword), listeners were biased to interpret the ambiguous sound as /f/, also subsequently in pseudowords, showing that they used lexical information to recalibrate speech categories [3], [4].

Another example is so-called audiovisual recalibration. This refers to listeners using visual (e.g. lip-reading) information to recalibrate speech perception. In these studies, listeners are exposed to videos of a speaker pronouncing a series of pseudowords. While the audio included an ambiguous consonant (for example, /a?a/, ambiguous between /aba/ and /ada/), the video was unambiguous in showing the speaker articulating either /aba/ or /ada/. Thus the visual information biased towards one of the two possible interpretations (for example, visual lip closure would bias towards /aba/). The results suggested that listeners had recalibrated their speech categories in a subsequent audio-only labeling task. Therefore, audiovisual information also may lead to recalibration of speech categories [5], [6]. A recent study compared lexical and audiovisual recalibration [7].

The present study adds to the literature by investigating audiovisual recalibration in vowel categories. While all studies mentioned above have focused on audiovisual recalibration of consonant categories, this study investigates whether similar results hold for vowel categories. With respect to lexically-guided recalibration, recent studies have already shown that it occurs with vowel categories [8]–[10]. In the present paper, we investigate whether this holds for audiovisual recalibration as well. In addition, it is well established that vowel categories are less sharply defined and may be less stable compared to consonant categories [11]. One may wonder whether the stability of phoneme categories (i.e., the "sharpness" of the phoneme boundaries) may affect this recalibration. If so, this may lead to two contrastive hypotheses. If the phoneme boundary is less sharp, this means the category is less clearly defined, and hence it could be more open to moving around. This would suggest listeners with less sharp boundaries show stronger recalibration effects. On the other hand, one could argue that if one has a sharp boundary and there is visual information that the boundary needs to move, one may be more likely to move it. This then would lead to stronger recalibration for listeners with sharper boundaries.

2. Methods

2.1. Participants

10 native Dutch participants (7 females; age: M = 23 (SD = 4.06)) were recruited and provided informed consent according to the declaration of Helsinki. Participants were randomly assigned to one of two participant groups (5 participants in each group).

http://dx.doi.org/10.21437/Interspeech.2017-122

2.2. Materials

A 22-step vowel continuum was created using Praat [12]. An original recording of the Dutch vowel /e:/ (spoken in context as /kapek/) was chosen and its source signal was extracted using linear predictive coding (LPC) and inverse filtering. The filter was manipulated by decreasing both F2 and F3 in 22 steps. The filter was then recombined with the source signal. The resulting vowels were recombined with the phonological context /kap_k/, resulting in a /kapek/-/kapøk/ continuum. So the whole continuum was created by manipulating F2 and F3 from a single /kapek/ recording. Both endpoints of the continuum are nonwords in Dutch.

Video stimuli were created by pairing each step of the audio vowel continuum with a video of the speaker's mouth articulating either /kapek/ (where the critical second vowel, /e/, is unrounded) or /kapøk/ (where the critical second vowel, /ø/, is rounded and hence visually distinct from /e/). This was the same speaker as in the auditory stimuli. Catch trial videos were created by adding a white dot (appearing for one frame only) in the middle of the video.

2.3. Paradigm

After instructions, participants were presented with a calibration block (audio only stimuli), a pre-test block (audio only stimuli), and then three to six (5 participants in each case) cycles alternating between exposure blocks (audiovisual stimuli) and post-test blocks (audio only stimuli).

In the calibration block, 12 steps along the continuum were presented (each 10 times) with a randomized order in a 2-alternative forced choice (2AFC) classification task. Participants were required to classify the stimulus as either /kapek/ or /kapøk/ by button press. For every participant individually, the most ambiguous step on the continuum (step x) and two neighboring steps (step x-2, x+2) were used in the remainder of the experiment. Step x-2 is closer to the /e/, and step x+2 is closer to /ø/.

In the pre-test, the same 2AFC classification task was used, this time including only the three most ambiguous steps (x, x-2, x+2). Each was presented 20 times, in randomized order.

In each exposure block, 20 videos were presented in a between-participant design. Videos were selected for each participant based on the participant group and their calibration phase data. For the /e/ group, videos of /e/ articulation (/kapek/) were paired with the audio of the most ambiguous step (step x), while videos of an /ø/ articulation (/kapøk/) were paired with an unambiguous /ø/ audio (step 22). In the /ø/ group, the /ø/ video was paired with the ambiguous audio (step x), while the /e/ video was paired with unambiguous /e/ audio (step 1). In every 20-trial block, two videos were catch videos (with a white dot). Participants were instructed to press a button as soon as they detected the white dot, in order to make sure they were looking at the videos.

The post-tests were similar in stimuli and task to the pre-test, but consisted only of 6 trials each (or 12 for half of the participants, but only the first 6 were analyzed). Exposure and post-test blocks alternated.

2.4. Analysis

All analyses were performed in R [13]. For the calibration phase, a logistic regression was applied to determine the most ambiguous step along the continuum for every participant. This was defined as the step closest to the 50% cut-off of the logistic regression curve (i.e., the point along the continuum that would be classified by the participant as /ø/ in 50% of the cases). This step and two nearby steps (steps x, x-2, x+2) were used in the other blocks of the experiment.

For the pre-test and the post-tests, a generalized (binomial) linear mixed model was fitted to the data using a Laplace approximation with the R 'lme4' package [14]. Post-hoc investigation of the interaction term was performed using Holm's method for multiple comparison correction.

Finally, in order to take a closer look at individual variability, we compared the steepness of each participant's logistic curve fitted to their calibration phase data (i.e., the sharpness of the phonemic category boundary) with that participant's learning effect. The latter was quantified as the absolute value of the by-participant random slope coefficients for Time (pre- vs. post-test) from the generalized mixed model fitted to the data from pre- and post-tests.

3. Results

Figure 1 shows the results over participants in the calibration block. The most ambiguous steps ranged across participants from step 6 to step 10. Based on the logistic regression, a logistic curve was fitted for each participant, as shown in figure 1. These logistic curves vary across participants by two parameters, describing the boundary location (the point where the curve crosses the 0.50 line) and the boundary sharpness (the steepness of the curve).

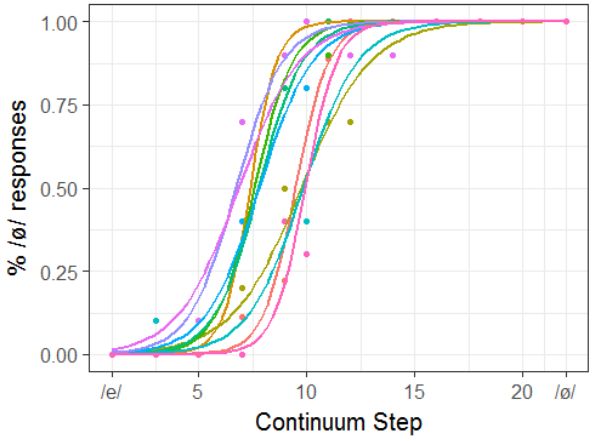

Figure 1: *Calibration results across participants. Colors represent participants; lines represent logistic regression fitted curves.*

Subsequently, the average percentages of /ø/ responses were calculated as a function of stimulus, test time (pre-test vs. post-test) and participant group. Figure 2 shows the results. Overall, the graph shows that most /ø/ percentages for the /e/ group (top row, /e/ group) decrease from pre-test to post-test, whereas this is not the case for the /ø/ group (bottom row, /ø/ group). The clearest effects can be seen for stimulus step x (i.e., the most ambiguous vowel for each participant).

In order to explore the effects of the exposure to the video clips, we fitted a generalized mixed effects model to the data with fixed effects Time (pre vs. post), Group, and Stimulus, as well as their interaction terms, and a random intercept for participants and by-participants random slopes for Time.

Analysis of the fixed effects estimates of the model shows significant main effects of Time (z = -2.87, p = .004), Group (z = 2.05, p = .040) and Stimulus (z = 8.48, p < .001), as well as a significant interaction between Group and Time (z = 2.26, p = .024). This suggests that the training affected the two participant groups differently. A closer analysis of the interaction effect revealed that while there was no significant difference between the pre-test and post-test for the /ø/ group ($\chi^2(1)$ = .23, p = .63), the /e/ group did categorize stimuli significantly less often as /ø/ in the post-test compared to the pre-test ($\chi^2(1)$ = 8.24, p = .0082), as was predicted.

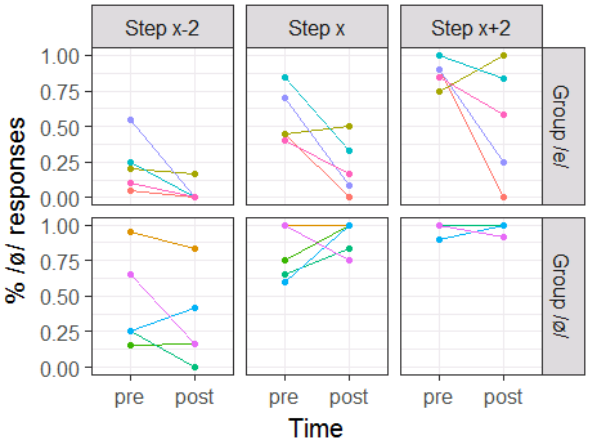

Figure 2. *Percentage /ø/ responses as a function of participant, participant group, stimulus step, and test time.*

The secondary aim of this study was to look at whether boundary sharpness was associated with the amount of audiovisual recalibration. Two alternative hypotheses were put forward. On the one hand, less sharp boundaries could be freer to move around and therefore show more susceptibility to recalibration. On the other hand, less sharp boundaries might stay fuzzy under adaptation conditions and sharper boundaries might be more prone to recalibration. The boundary sharpness for each participant (steepness of the curves in figure 1) tended to be associated with the participant's learning effect.

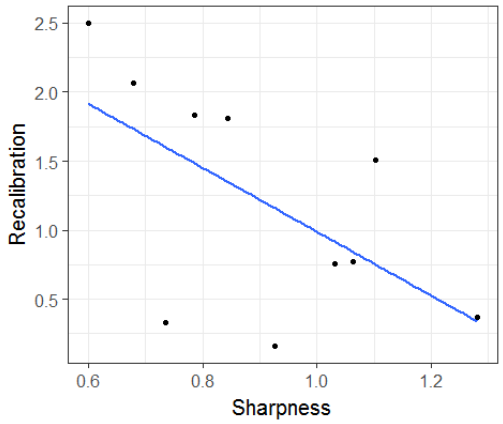

Figure 3: *Individual amount of audiovisual recalibration as a function of phoneme boundary sharpness.*

Specifically, as the boundary steepness decreased, the learning effect increased (Figure 3), which is in line with the first hypothesis. However, caution is warranted for this interpretation, given this association was not significant (Pearson's $r(8)$ = -.60, p = .069; Spearman's $\rho(8)$ = -.53, p = .12).

4. Discussion

In this study, we investigated whether listeners would recalibrate vowel categories using audiovisual information. The results suggest that this is indeed the case: The audiovisual perceptual learning block led to different effects in the two participant groups. Specifically, the group that was exposed to /e/ videos paired with ambiguous audio showed a reduction in /ø/ responses after perceptual learning.

These results suggest that participants recalibrated their vowel categories by shifting the /e/-/ø/ phoneme boundary. Recalibration of phoneme categories is one way by which listeners can attempt to solve the mapping problem between the incoming acoustic signal and abstract phonological categories. This is in line with what previous studies showed with audiovisual recalibration in consonant categories [5], with lexically-guided recalibration in vowel categories [8]–[10], and with the broader literature of what is known as cross-modal recalibration, for example in spatial cognition [15].

The current data do not make it possible to investigate the temporal development of audiovisual recalibration. Previous studies on audiovisual recalibration in consonants have suggested that recalibration is a short-lived effect. A comparison between audiovisual and lexical recalibration suggested that audiovisual recalibration effects lived only for up to five test tokens after exposure [7]. The current study does not allow us to see whether this also holds for vowels, or whether the effect would last longer or shorter, given the less sharply defined phonemic boundaries in vowels than in consonants.

Closer investigation of the interaction between Time and Group suggested that only the /e/ group showed a difference between pre- and post-tests. The lack of audiovisual recalibration in the /ø/ group was unexpected. Inspection of individual participants' results showed that there was quite some variability across participants, also within the /ø/ group. Interestingly, the results in figure 2 show that for step x in the /ø/ group, three out of five participants did show a change from pre- to post-test in the expected direction (an increase), while the two participants that did not show this effect already were at 100% in the pre-test (so they could not have shown any increase). This suggests that the lack of a group-level recalibration for the /ø/ group is a ceiling effect: there simply was no way to show recalibration. The 100% /ø/ responses in the pre-test suggest that the calibration block at least for these two participants did not adequately estimate the phoneme boundary. Moreover, also the other participants tended to show pre-test % /ø/ responses of over 50%, suggesting that step x may have been less ambiguous than was thought. One reason for this may be the asymmetry in the calibration stimulus materials. From figure 1, it can be seen that all participants had their boundary in the left half of the continuum, and none of them categorized one of the four most /ø/-like stimuli (steps 16, 18, 20, 22) less than 100% as /ø/. This asymmetry may lead to a bias in participants' response patterns, as unbalanced exposure to two categories (e.g., more exposure to /ø/ stimuli compared to /e/) may bias subsequent

categorization of these categories (as in selective adaptation [16], [17]) and therefore a mis-estimation of the boundary location. Future research should try to control for this bias, for example by excluding stimuli that don't differentiate between participants.

Finally, a closer analysis of individual variability in the data suggested that the amount of recalibration may be associated with phoneme boundary sharpness (although this result definitely needs replication). Specifically, the present findings indicate that well-defined categories are more robust to audiovisual recalibration, whereas less sharply defined categories are more susceptible to be recalibrated through audiovisual integration. Assuming this result shows up more robustly in a better-powered study, this suggests that listeners with fuzzy category boundaries assign more weight to visual information during audiovisual speech recognition. This may be explained as participants with less well-defined boundaries may find stimuli straddling the category boundary to be more ambiguous and therefore these participants stand to gain more from visual information compared to the participants with sharper category boundaries.

5. Conclusions

This study shows that listeners use visual information to recalibrate their vowel categories. This is in line with past research on consonants and lexically-guided recalibration, but extends it to vowel categories, which are known to have less sharply defined categorical boundaries. Moreover, although the current data do not warrant any strong conclusions about individual differences, it was suggested that individuals with fuzzy or less sharp perceptual category boundaries assign more weight to visual information during audiovisual speech recognition and therefore show increased audiovisual recalibration. If this finding is corroborated, given that vowel categories have less sharp boundaries compared to consonants, there ought to be audiovisual recalibration for vowel categories, given consonants have shown audiovisual recalibration in previous research. This is indeed what was found in the current study.

6. References

[1] T. M. Nearey, "Static, dynamic, and relational properties in vowel perception," *J. Acoust. Soc. Am.*, vol. 85, no. 5, p. 2088, 1989.

[2] W. F. Ganong and W. F., "Phonetic categorization in auditory word perception.," *J. Exp. Psychol. Hum. Percept. Perform.*, vol. 6, no. 1, pp. 110–125, 1980.

[3] D. Norris, J. M. McQueen, and A. Cutler, "Perceptual learning in speech," *Cogn. Psychol.*, vol. 47, no. 2, pp. 204–238, Sep. 2003.

[4] F. Eisner and J. M. McQueen, "The specificity of perceptual learning in speech processing," *Percept. Psychophys.*, vol. 67, no. 2, pp. 224–238, Feb. 2005.

[5] P. Bertelson, J. Vroomen, and B. De Gelder, "Visual recalibration of auditory speech identification: a McGurk aftereffect.," *Psychol. Sci.*, vol. 14, no. 6, pp. 592–7, Nov. 2003.

[6] A. Ley, J. Vroomen, L. Hausfeld, G. Valente, P. De Weerd, and E. Formisano, "Learning of New Sound Categories Shapes Neural Response Patterns in Human Auditory Cortex," *J. Neurosci.*, vol. 32, no. 38, pp. 13273–13280, Sep. 2012.

[7] S. van Linden and J. Vroomen, "Recalibration of phonetic categories by lipread speech versus lexical information.," *J. Exp. Psychol. Hum. Percept. Perform.*, vol. 33, no. 6, pp. 1483–1494, 2007.

[8] J. M. McQueen and H. Mitterer, "Lexically-driven perceptual adjustments of vowel categories," *ISCA Work. Plast. Speech Percept.*, no. June, pp. 233–236, 2005.

[9] K. Chládková, V. J. Podlipský, and A. Chionidou, "Perceptual adaptation of vowels generalizes across the phonology and does not require local context.," *J. Exp. Psychol. Hum. Percept. Perform.*, vol. 43, no. 2, pp. 414–427, 2017.

[10] J. Maye, R. N. Aslin, and M. K. Tanenhaus, "The weckud wetch of the wast: lexical adaptation to a novel accent.," *Cogn. Sci.*, vol. 32, no. 3, pp. 543–562, 2008.

[11] P. K. Kuhl, "Human adults and human infants show a 'perceptual magnet effect' for the prototypes of speech categories, monkeys do not," *Percept. Psychophys.*, vol. 50, no. 2, pp. 93–107, Mar. 1991.

[12] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer Program]." 2013.

[13] R Core Team, "R: a language and environment for statistical computing." R Foundation for Statictical Computing, Vienna, Austria, 2013.

[14] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *J. Stat. Softw.*, vol. 67, no. 1, 2015.

[15] F. L. Bedford, "Constraints on perceptual learning: objects and dimensions," *Cognition*, vol. 54, no. 3, pp. 253–297, Mar. 1995.

[16] D. F. Kleinschmidt and T. F. Jaeger, "Re-examining selective adaptation: Fatiguing feature detectors, or distributional learning?," *Psychon. Bull. Rev.*, vol. 23, no. 3, pp. 678–691, Jun. 2016.

[17] P. D. Eimas and J. D. Corbit, "Selective adaptation of linguistic feature detectors," *Cogn. Psychol.*, vol. 4, no. 1, pp. 99–109, Jan. 1973.